Pipelines de ETL e Machine Learning com Apache Spark

Projeto 5
Otimização de Pipeline ETL e Machine Learning com PySpark

Neste projeto nosso objetivo é construir um pipeline otimizado, tanto no nível do cluster Spark, quanto no nível do job de processamento (código Python), a fim de validar 3 diferentes técnicas de pré-processamento de dados de texto. O resultado de cada técnica será usado para construir e avaliar um modelo de Machine Learning e assim selecionar a técnica que permite construir o modelo com melhor performance. Tudo isso executado em um multi-node cluster Spark.

Os dados usados no Projeto foram preparados com base nos dados disponíveis no link abaixo:

https://ai.stanford.edu/~amaas/data/sentiment/

Bons estudos.

Equipe DSA

Muito Obrigado!
Continue Trilhando Uma Excelente Jornada de Aprendizagem.